Team Control Number

For office use only
T1 _____
T2 _____
T3 _____
T4 _____

30407

Problem Chosen

C

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2014 Mathematical Contest in Modeling (MCM) Summary Sheet
(Attach a copy of this page to your solution paper.)

Type a summary of your results on this page. Do not include the name of your school, advisor, or team members on this page.

Using Networks to Measure Influence and Impact

Networks in the real world can be classified into two categories, namely undirected ones and directed ones. In this paper, we propose two important measurements in our model to determine influential nodes in both types of networks. One measurement named by Modified PageRank (MPR) is based on theories of eigencentrality and Google's PageRank algorithm, which applies to both undirected and directed networks. Another measurement called Accumulated Citation Algorithm (ACA) is developed from Breadth First Search (BFS) algorithm. It can only be used to analyze the directed networks with tree structure.

To develop MPR, a series of concepts are introduced from law of gravity in physics including "virtual distance", "force" and "acceleration". These newly-defined concepts work well in quantifying the influence of nodes and help to extend the range of application of our model.

Based on our measurements, the three most influential authors within Erdos1 Network are identified to be WILSON RICHARD MICHAEL, CONWAY JOHN HORTON and FELLER WILLI K. (WILLIAM)*. After the discussion of assigned tasks, sensitivity analysis is conducted to examine the stability of our model. The testing results indicate that our model can resist small disturbance from outside. Finally, we recognize the strengths and weaknesses of our model and present some approaches to refine our model.

Key words:
Modified PageRank, Eigencentrality, Accumulated Citation, virtual distance, force, acceleration

I.Introduction

Network science is a very young subject with lots of interesting findings. Networks are ubiquitous. From small scale such as family and school to large scale like nation and the Internet, we actually live in a world full of networks. Without doubt networks are important to human beings, however, real networks are usually too complicated to analyze, not along to have any understanding about them. Therefore, one of the missions of network science is to develop valid way to model real networks.

One of the most important issues in a real network is to measure the influence and importance of a member in this network. So far, there have been quite a few criteria to measure and quantify the importance of a member. The most common way to find important components is to compute the centrality [1]. There are many different kinds of centrality including degree centrality, closeness centrality, betweeness centrality and eigen centrality [1, 2]. In this paper, we will mainly focus on eigen centrality because its huge potential in many area and successful applications. One of the most famous examples is the PageRank algorithm invented by the Google funder Larry Page.

Generally speaking, PageRank is just one of the variations of eigen centrality. As many other centrality indices arose from the studies of complex networks, eigen centrality aims to characterize the influence of the members in a network. However, while many of the methods focus on how to capture the properties of a specific individual in the network, eigen centrality tries to consider the interaction of all the elements in a network simultaneously [3]. It not only considers the neighborhoods but also considers the neighbor of neighborhoods.

The basic form of eigen centrality starts from expressing the quantity of an individual by the following equation revealing the relationship of it and other individuals in a network:

$$x_i = \frac{1}{\lambda}\sum_{j=1}^{n} \mathrm{A}_{ij}\, x_j$$

Where x_i is the quantity of the i-th vertex in the network, A_{ij} is the entry of the adjacency matrix of the network and λ is a constant. If we define the vector of the values of vertices to be $\mathbf{x} = (x_1, x_2 \ldots x_n\ )^T$, then a matrix equation can be formed as

$$\mathrm{A}\mathbf{x} = \lambda\mathbf{x}$$

It is exactly the eigen problem formulation and λ and $\mathbf{x}$ are eigenvalue and eigenvector respectively. Once the eigenvector is calculated, it can be used to form the measurement of the impact or other quantities like Google does.

In the PageRank method it is also necessary to solve eigen problems, but there is a little difference. PageRank utilizes the features of the Internet and assigns probability to the edges. A user recently staying at a vertex have probability p to link to other Webs to which this vertex points and have probability $(1-\mathrm{p})$ to randomly surf around the Internet. Besides, if there are d links (with the same weight) pointing out, it is natural to assume that the user have $1/d$ probability to choose one of the links. Based on the above, the transition matrix can be written as described in [4]

$$M_{ij} = \begin{cases} \frac{pa_{ij}}{c_j} + \delta, & if\ c_j \neq 0 \\ \frac{1}{n}, & if\ c_j = 0 \end{cases}$$

Here $c_j = \sum_i a_{ij}$ is the column sum, n is the number of all the vertices and δ is $(1-\mathrm{p})/\mathrm{n}$. Also note that in the formulations above the graphs are all directed graph and if $A_{ij} = 1$, it means that the jth node points to the ith node in the graph. We notice that the column sums of M are always equal to 1 and hence it is a Markov transition matrix. An importance result from matrix theory known as the Perron-Frobenius theorem ensures that such a matrix can have an unique eigenvalue equals to 1 [5]. In other words, the equation $\mathrm{M}\mathbf{x} = \mathbf{x}$ can be solved and the corresponding eigen vector is just the PageRank used by Google!

In this report, we calculated the importance based on the PageRank method but do not use it directly because the networks under discussion don't possess much random phenomena like the WWW. To make an analogy to the WWW network, we assume there is a "transmission of importance" between any two linked nodes in a network. Like internet users may tend to visit big websites, the importance may "transmit" from some places to bigger ones or more important ones. For example, if you know a poor guy and rich guy (that is, you are linked with them), you may probable be more active with the rich guy. In some sense, we can say that the importance of the small three people network is transmitted to the rich guy.

After introducing the concept of importance transmission, Google's PageRank algorithm is modified to be a new algorithm called Modified PageRank (MPR) to fit in the problems. Besides, we propose many new ideas with corresponding quantities to help apply the MPR algorithm to networks which are significantly different from the Internet. The outline of the remaining parts is as followed. In ch2 the basic assumption used through this report will be presented. In ch3 we will describe the prototype of our model and characterize the mathematical formulations. Ch4 will address Task1 and Task2. From ch5 to ch7 we will deal with Task3 through Task5 consecutively. In ch8 sensitivity analysis will be performed and in ch9 the strength and weakness of our model will be discussed.

II. Assumptions and Notations

2.1 Assumptions

We make the following assumptions about computing the influence ranking of nodes in networks.

1. We do not take into account the differences in format, length of original text and position in papers between citations. And in fact the prestige of citers affect the importance of citations, but in order to sim`e5plify the model, the difference in prestige of citers between citations is neglected.

2. Each time of cooperation with Paul Erdos has the same contribution to the indirect connection which is measured by a variable called "virtual distance" between every two authors, regardless of details of co-authorship. Similarly, we assume that each time of cooperation between every two authors makes the same contribution to the direct connection which is measured by a variable called "actual distance" between them.

3. The influence of difference in publication year between papers is neglected.

4. We assume that for each paper, every author makes the same proportion of contribution to the paper. Similarly, every leading role in the movie makes the same proportion of contribution to the movie and this assumption also applies to supporting roles.

2.2 Definitions of Symbols

$\deg(x_i)$	Degree centrality of author i, measured by the number of co-authors and the prestige of co-authors, it is the sum of the prestige of each co-author multiply the weight of the connection path between the author and his co-author [1, 2].
$\mathrm{clo}(x_i)$	Closeness centrality of author i, measured by the sum of the inverse of the shortest path between author i and author j (j≠ i) [1, 2]
$\mathrm{bet}(x_i)$	Betweenness centrality of author i, it is the sum of total number of shortest paths between each two different authors (except author i) divided by number of shortest paths between these two different authors passing through author i [1, 2].
λ	Eigenvalue
$\mathbf{x}$	Eigenvector

III.Model Establishment

Transition of importance is very abstract and lack of obvious physical meanings and hence it is not intuitive to construct a transition matrix for such networks. Nevertheless, once the physical meaning is clarified, the transition matrix can be build and the MPR can easily be calculated through eigen vector calculations. Instead of using the

traditional probabilistic formulations, the concept of "virtual distance", "combined distance" and "node mass" are developed to help build the transition matrix and calculate the MPR. However, those special quantities are only used in undirected graphs, not directed graphs. It's due to the properties of those two kinds of networks. For instance, it's very hard to construct the concept of distance in a directed graph. Therefore, it is necessary to distinguish our different approaches towards undirected and directed graphs.

3.1 Undirected Graph

In this paper, networks built through cooperation are treated as undirected graph. An natural explanation is that when two individuals work together, they can both access the resource obtaining by each other and hence it is reasonable to model those networks as undirected graphs. We define the transition of importance in such networks to be the "force" drives an individual to cooperate with others. The higher the force is, the more the importance is transmitted to the other side. But how to calculate the force? We borrow from physics the famous law of gravity, using the special quantities mentioned above to compute the unnormalized value of M_{ij}.

To do this first we have to find the combined distance $d_c(i,j)$. $d_c(i,j)$. is computed as follows:

$$d_c(i,j) = \frac{1}{\frac{1}{d_a(i,j)} + \frac{1}{d_v(i,j)}}$$

Where $d_a(i,j)$ is calculated by the unweighted shortest path and if the value is Inf, we just ignore this term. The MATLAB toolbox provided by [6] can easily find the all shortest path matrix. $d_v(i,j)$ is our defined quantity representing the closeness within two nodes in a network. To find this, first we have to identify a common quantity shared by all individuals in a graph (whether this graph is connected or unconnected). For example, the people in Erdos1 network all have co-authored with Erdos. We define the value of this common property of the i-th node as com_i and the virtual distance is

$$d_v(i,j) = \max\{\{d_a(i,j) | i,j \in V\} \backslash Inf\} + com_i * com_j$$

The reason why adding the maximum reachable distance of the graph is that we don't want the influence of virtual distance exceed the actual distance. Also note that $d_c(i,j)$ is always nonzero and this means that the new connected matrix D_c is a complete graph. This bring the advantage when doing numerical calculation because 0 may introduce numerical instability and now the problem is eliminated.

Another important quantity is the mass. We define the mass of a node to be the strength or competence in terms of total quantity and average quantity. These information cannot

be known simply by the internal physical structure of a network. For example, the total citation and average citation can used to determine the mass, of an author in an academic network. The formula for the mass of the i-th vertex is defined below:

$$m_i = s_{avg}(v_i) + 5 * \ln(s_{total}(v_i) + e)$$

Here s_{avg} means the average quantity and s_{total} means the total quantity. The additional natural constant e is used to avoid negative value. With those quantities, now it's time to form the equation for the network force.

$$F_{ij} = \frac{m_i * m_j}{d_c(i,j)^2}$$

After F_{ij} is computed, the unweighted transition matrix is just the force matrix F. And then normalizing it, calculating the eigen vector of M with corresponding eigen value 1(which is uniquely the largest eigen value). Finally this eigen vector is just the MPR. The procedures and flow chats for the modeling of undirected graphs are presented below:

(1) Build the connected matrix A of the network and collected necessary data.
(2) Calculate combined distance $d_c(i,j)$ through $d_a(i,j)$ and $d_v(i,j)$.
(3) Calculate mass m_i for each node according to the data.
(4) Compute the force F_{ij} between every two nodes
(5) Form the transition Matrix M by normalizing F
(6) Find the eigen vector **x** of M with eigen value 1. It is the MPR for undirected graphs.

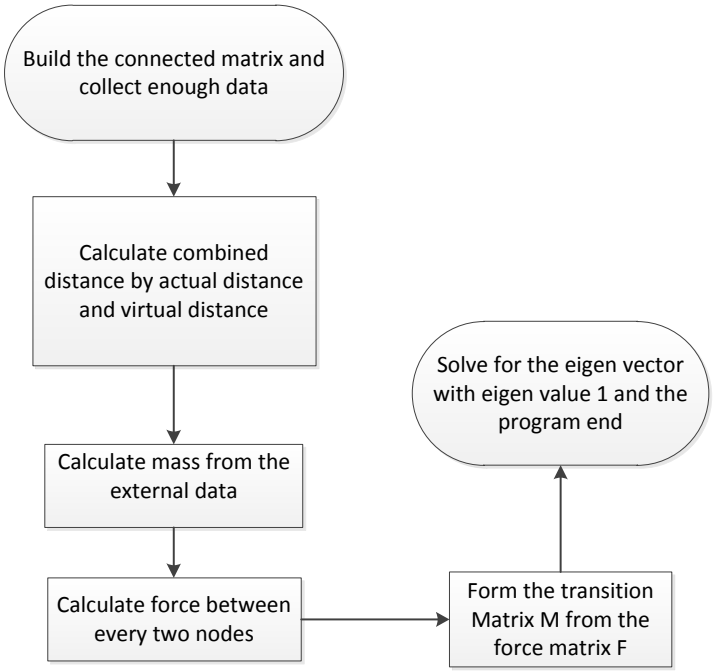

Fig1. The algorithm for undirected graphs

3.2 Directed graph

There are two algorithms in this section. The first one is still the MPR algorithm.

In a directed graph like the one appears in Question 3, if node i points to node j, it means that i needs something from j. In this situation it is assumed that j has more impact that i and importance can only be transmitted from i to j. Hence the whole situation is like the Internet and the MPR is without too much modification. There are only two differences. First, the adjacency matrix A is weighted by external data before normalization. How to calculate the weighted value depends on the characteristic of the network and hence we won't discuss here. Second, if the sum of a column of A , say column i, is zero(this means that you can go nowhere from this vertex), then do the following:

(1) Assign a high probability, say p = 0.9, to the diagonal term A_{ii}.
(2) Add every terms of this column a value (1-p)/n, where n is the matrix dimension.

The physical meaning of this is that when the vector $\mathbf{x}$ is been transmitted, the quantity of the i-th node x_i will mostly remain the same since there is nowhere for it to transmit! You may argue that why not set the diagonal term to 1? However, it introduce many zeroes to other entries hence we do not consider this. The remaining steps is the same as MPR.

Other than MPR, the other algorithm which we called "Accumulated Citation Algorithm"(ACA) is proposed to handle the secondary importance transmission. It can only be used in Tree structure. The basic idea is that if A points to B and B points to C, then part of the importance will transmitted from A to C. But since A is more distant than C, the quantity transmitted from A to C will less than those transmitted from A to B. In this report a decay factor b will be used. The algorithm is basically based on BFS(Breadth First Search) and an impact quantity a_i representing the importance of the i-th node. The procedure is as follows.

(1) First calculate the level of every node. It can be done by repeatedly looking for leaves of the tree (without any element pointing to it) and record them.
(2) For each node in the same layer, use BFS to visit every reachable nodes
(3) Add $a_i * b\hat{}(c)$ to the visited nodes. Where c is the layer difference between the i-th nodes and the nodes it visited. The quantity needs to be calculated by decay factor.
(4) If we haven't investigated all the layers, layer = layer + 1 and go back to step(2), else continue this procedure.
(5) Compute the Compute the total cumulated value of a node and it is the importance quantity of that node
(6) The detail methods used is described in the flow chart below. For simplification, we only present the flow chat of ACA .

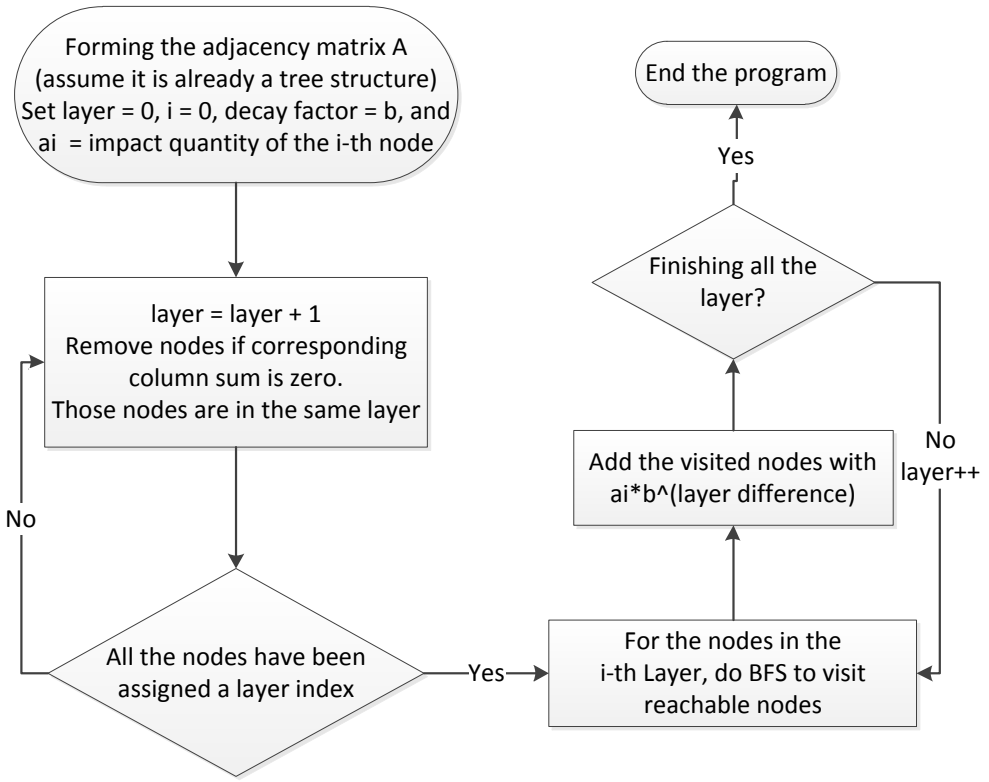

Fig2

IV. Task 1 & 2

There are two columns of original information. As stated by the task, the first column is a list of the 511 coauthors of Paul Erdos and the second column specifies coauthors of every member in the list of Erdos1.

In order to illustrate the co-authorship within the 511-person network, a symmetric matrix P with 511 rows and 511 columns is built. If author k and author j have cooperated before, then P_{ij} and P_{ji} are assigned with 1. Otherwise, the values of P_{ij} and P_{ji} are both 0. Collaboration with researchers outside the Erdos1 Network is irrelevant to our analysis and therefore omitted.

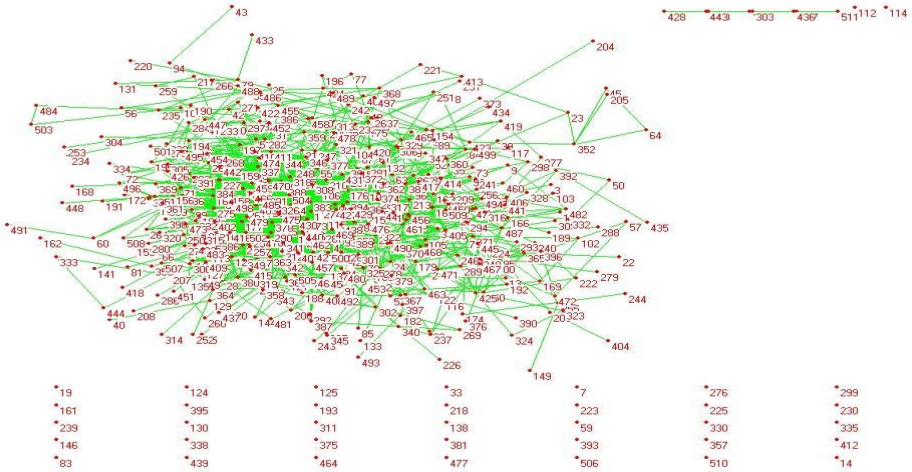

Fig3

The constructed network is shown above.

We concentrate on the several key properties of the network, namely degree, closeness and betweenness. Histograms are depicted to show our results in an intuitive way.

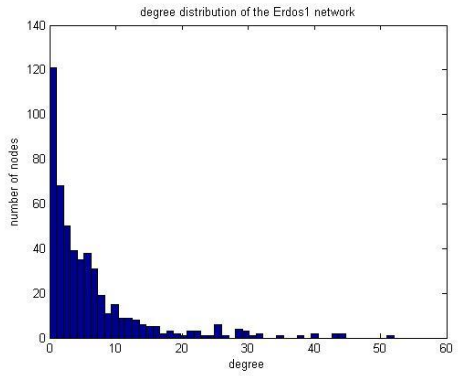

Fig4

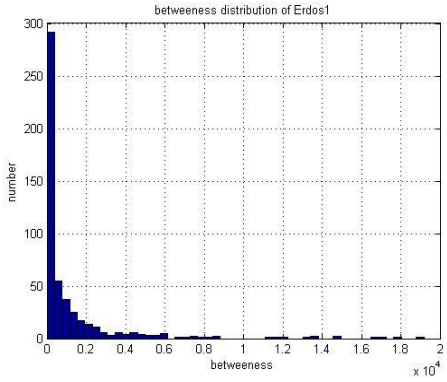

Fig5

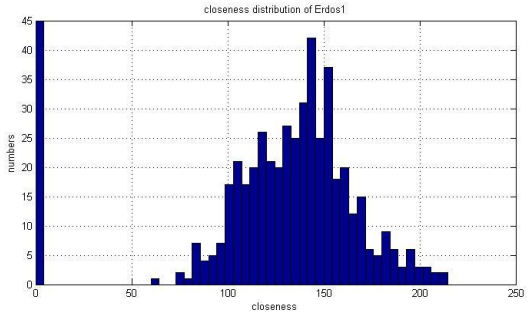

Fig6

Firstly, in terms of degree distribution (Fig4), we observe that a majority of nodes in the network have a degree centrality less than 10, which reveals insufficient interconnection and lack of complexity within the network. In contrast, we also notice that there exist several nodes of which the degree centrality exceeds 40, indicating possible significant power within the network.

Secondly, it is not surprising that betweenness graph (Fig5) is very similar to that of degree distribution graph since both of these two attributes describe the extent of significance in the network.

Thirdly, note that closeness computed here is slightly different from the generally accepted definition, which is the reciprocal of farness. This correction is made to accommodate the need to consider the existence of disconnected components. An approximately normal distribution is shown (Fig6). This result is consistent with finding in previous parts since only a few nodes appear to have good connections in the network.
In addition, based on our calculation, number of the isolated nodes is 37, which is less than one tenth of the total number.

With respect to Task 2, it is obvious that co-authorship has no direction and therefore

Erdos1 Network should be classified into undirected graph. In order to apply the algorithm derived in section I of chapter III, external data of total citation and average citation of all the 511 authors are collected from *American Mathematical Society* [7] to compute their mass. Eventually, we can obtain the ranking information of the network. Top 10 influential authors are shown in the following table.

Name	MPR	Total Citation	Average citation
WILSON, RICHARD MICHAEL	0.006090213	1524	24.98360656
CONWAY, JOHN HORTON	0.005689597	4451	26.33727811
FELLER, WILLI K. (WILLIAM)*	0.005219018	3908	41.57446809
THOMASSEN, CARSTEN	0.005110836	2160	10.33492823
JANSON, SVANTE	0.004957845	2951	11.05243446
DIACONIS, PERSI W.	0.004908215	3234	16.00990099
WORMALD, NICHOLAS CHARLES	0.004882831	1356	6.848484848
GODSIL, CHRISTOPHER DAVID	0.004872149	1662	16.13592233
MONTGOMERY, HUGH LOWELL	0.004861979	1620	18.62068966
HELL, PAVOL	0.004809186	1632	8.967032967

According to our assumptions and definitions, when ranking authors, both of their total citations and average citations should be taken into account. In other words, authors who rank high in the list must have relatively higher total citations and higher average citations. Using the MPR algorithm, the result we obtain quite matches our estimation. Therefore, we can declare the MPR algorithm successfully rank the authors according to their influence.

For comparison purpose, scatter plot of degree centrality of every node versus its rank is illustrated in the following graph.

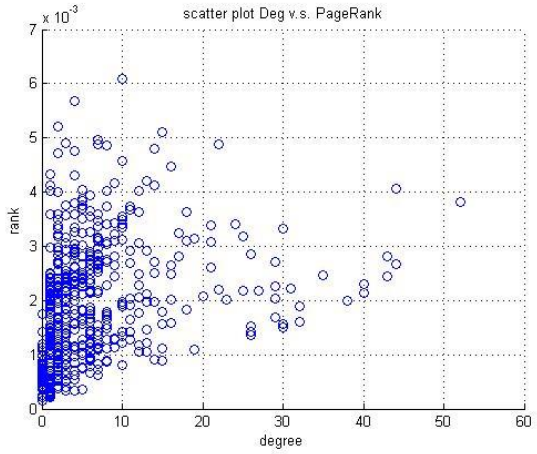

Fig7

We can see from the plot that clustering tendency exists in the interval of 0 to 10 degree of centrality. This can be attributed to the following reason. Some authors may not collaborate with many other researchers, resulting in their low degree of centrality. However, this does not necessarily indicate low quality of their published works. As verified by external data, they may also have desirable rankings.

V. Task 3

Note that due to the fixed sequence of publication year of different papers, no inter-citation exists, which indicates the constructed network is actually a spanning tree with directed paths. Therefore, algorithms developed in section 2 of Chapter III applies. In order to fulfill the requirement of Task Three, 21 papers including those attached to the problem and additional ones retrieved from Google Scholar [8] are collected for analysis. Detailed citation relationship among 21papers is reflected by the following Figure.

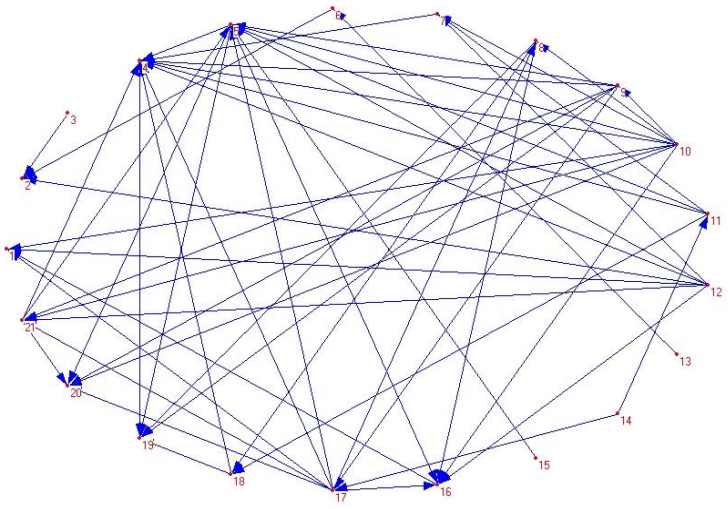

Fig8

5.1 MPR Algorithm

The implementation of weighting in this task is based on the total citation of the papers. If A is cited x times (i.e., with x total citations) and A points to B (B is cited by A), then the weighting of B will be added by ln(x). After computing all the weights, the only remaining task is to assign probability and calculate MPR. And in this task, dimension of transition is 21 (n=21).

5.2 Accumulated Citation Algorithm

Basically, the aggregate times a thesis is referred to represents its academic influence (importance). However, such method neglects the existence of “Mentor Effect”. The terminology “Mentor Effect” is used to qualitatively describe the positive impact of a previous cited article on the current thesis being discussed. Hence, it is reasonable to argue that citation of a mentee’s paper can be partly attributed to the help from mentor,

and thus should be counted into the calculation of a mentor's influential power. To simplify the discussion, a coefficient of 0.1(b=0.1) is applied. More specifically, total citation figure of a specific article should be increased by the summation of one tenth of cited times of every subsequently published paper referring to it.

5.3 Comparison between ranking results of two algorithms

Paper No.	original total citation [8]	cumulative total citation	Rank I	MPR	Rank II
1	4531	5455.835102	8	0.181206	2
2	1956	2149.52	12	0.171042	3
3	498	498	19	0.004144	16
4	21688	23821.57242	1	0.026062	5
5	18843	19208.4341	2	0.015652	6
6	763	793.4	17	0.008287	8
7	1246	1581.76	14	0.006315	12
8	1523	1877.27	13	0.005405	14
9	2748	4073	9	0.00458	15
10	13250	13250	3	0.004144	16
11	835	903	16	0.006069	13
12	10616	10616	4	0.004144	16
13	304	304	20	0.004144	16
14	680	680	18	0.004144	16
15	33	33	21	0.004144	16
16	5278	7261.35108	6	0.01194	7
17	2317	2863.6	10	0.007041	9
18	929	1153.125	15	0.006911	10
19	5296	8103.311186	5	0.398762	1
20	3867	5860.72664	7	0.119158	4
21	2388	2806.866	11	0.006708	11

It can be seen from the table that papers No 1, 4, 5, 16, 19 and 20 are on the top of both lists. This result is in line with reality since *"On Random Graphs"(1), "On the evolution of random graphs"(16), "The small world problem"(19), "Collective dynamics of 'small-world' networks"(4), "Emergence of scaling in random networks"(5) and "Internet: Diameter of the world-wide web"(20)* are all pioneering works in network science.

However, the rankings of some authors vary largely. For example, authors No 10, 12, 9, 2, 6, ranks 3, 4, 9, 12, 17 respectively in The Accumulative Citation Algorithm while ranks 16, 16, 15, 16, 8 in The MPR Algorithm. Such difference can be explained by the different underlying mechanism of two algorithms. One significant disparity is that ACA focus on the present state of the network while MPR Algorithm concentrate more on the steady state of network in the future. It can be predicted if there is no disturbance

from outside, the ranking result of ACA will converge to that of MPR Algorithm.

5.4 Measurement of the influence of a specific university

The influence of a specific university is similar to the influence of a specific author, which means the model that has been established in task 2 can be calibrated and applied here. The simplest way to evaluate a university is to sum up the influence of all of its researchers. But this approach has a shortcoming, which is that researchers within a university would have a lot of cooperation. Therefore, the influence of their research findings would be calculated repeatedly and the influence of university would be overestimated. The influence of an author in one specific research finding should be separated from his co-authors. More precisely speaking, we should consider the author's contribution to his research findings more accurately, such as considering the author rank in the paper when calculating how much influence the times of citations of the paper can bring to the author.

Another factor that contributes to the influence of a university is the cooperation project it carries out with other universities. Similar to the co-authorship of an author with other authors, the number of cooperation projects of a university with other universities can be used to calculate the distance between different universities, and the more projects they co-found, the shorter distance between them is. Also, a university has its own mass. When calculating the mass of a university, more factors should be considered, such as the academic reputation, employer reputation, citations per faculty, international faculty and international students. After obtaining the mass of each two universities and the distance between them, we can calculate the "force" between two universities. With the force, we can calculate the eigenvalue of each university using the MPR algorithm. Eigenvalue reflects the ranking of universities, universities with higher ranking have bigger eigenvalue.

VI. Task 4

Partnership between movie stars is one of the most common relationships of cooperation. Different actors and actresses are connected into a tremendous network through starring in a variety of movies as either leading or supporting characters. Therefore, network in the filmdom is very similar to that in the academic field. Since there are no directions in the cooperation between performers, the network should be categorized as an undirected graph and algorithm in section I of Chapter III applies. In this task, films produced by renowned Universal Picture Co. in recent years are collected from Internet Movie Database (IMDB) [9] as a set of raw data. And for each selected movie, both leading and supporting performers are included in the construction of network. The three-dimensional Picture of the constructed network is shown as follows.

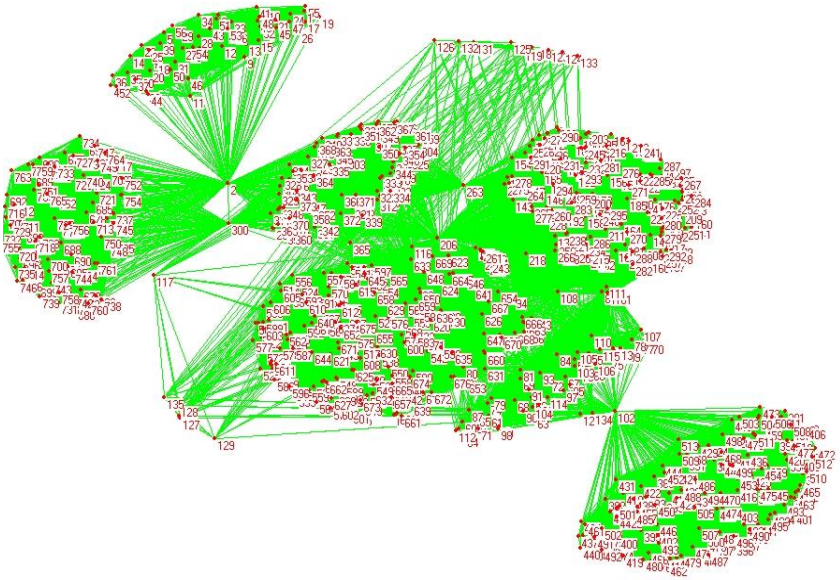

Fig9

The actual distance between two actors is defined by the shortest path connecting them. However, unlike Task 2, it is almost impossible to find a performer who has experiences in acting together with all the rest of his or her peers. In order to implement the algorithm derived before, Universal Picture Co. is assumed to play a similar role as Erdos in Task 2. In this way, virtual distance and combined distance between two performers A and B can be well-defined.

After that, detailed information of every chosen movie including ratings, number of wins and number of nominations is used to comprehensively evaluate its influential power. "Mass" of an actor is calculated by the summation of adjusted influential power of movies that he or she has ever played a role in. To recognize the contribution of leading actors to the success of a film, an adjusted coefficient of e (natural constant) is applied. And for the supporting characters, adjusted coefficient is simply 1.
Based on above adjustments, "force" between any two performers can be obtained and eigenvector of the transition matrix can be computed. Top 10 stars according to MPR are listed below. Degree of centrality are also calculated and listed for reference purpose.

Name	MPR	Name	Degree
'Asquith-Coe, Lee'	0.003120754	'Asquith-Coe, Lee'	411
'Choong, Siong Loong'	0.003072978	'Choong, Siong Loong'	383
'Chestnut, Morris'	0.002695263	'Ruben, Eddie'	289
'McAdams, Rachel (I)'	0.002617355	'Minelli, Morris'	267
'Nighy, Bill'	0.002617355	'Curry, Graham'	267
'Gleeson, Domhnall'	0.002617355	'Duggan, John (VI)'	267
'Duggan, John (VI)'	0.002465716	'Golt, David'	267
'Moore, Martyn'	0.002465716	'Harris, Lee Nicholas'	267
'Curry, Graham'	0.002465716	'Herdman, Richard'	267
'Golt, David'	0.002465716	'Liebman, Simone (I)'	267

Ironically, none of the top ten alleged "influential actors" based on our analysis has ever taken the leading role in any of the selected movies. This undesirable finding possibly comes from the fact that most of films chosen were released in year 2013. Hence, the huge network constructed may not be attributed to the leading actors because it is unrealistic for an actor to take the leading role in both films simultaneously. Conversely, those supporting characters may actually act as the intersection points that connect those relatively isolated networks.

Therefore, our model for undirected graph should not be suspected because of the failure to distinguish influential individuals from the network. It is the arbitrary selection of the movies that should be blamed. To verify the model in a more scientific manner, we contend that a set of films that were released in various points in time is more favorable.

VII. Task 5

Previous models and conclusions about network science have some implications on real world relationship between individuals, corporations or even countries. Now consider how individuals can use influence methodology to improve relationships. In order to find a co-author, there are two main factors to consider—the influence of the co-author and the convenience to cooperate with the co-author. Firstly, we use the influence methodology to compute the "mass" of all potential co-authors. Then, we calculate the accessibility (convenience) of all potential co-authors, which could be measured by "combined distance" defined in previous tasks. However, the model derived in preceding tasks cannot be transplanted here without refinement since we only consider about the impact of a prestigious author on a particular successor, not the other way round. In other words, only one of a pair of interactive forces is relevant to our discussion.

Inspired by Newton Second Law, we introduce the new concept of acceleration (denoted by a) here to quantitatively describe the speed of boosting reputation through co-authorship.

$$F_{ij} = \frac{m_i * m_j}{d_c(i,j)^2} = m_i * a$$

Where m_i and m_j are the mass of an individual and his or her potential co-author respectively

By cancelling the mass of current individual on both sides, we can obtain

$$a = \frac{m_j}{d_c(i,j)^2}$$

It can be observed that acceleration is positively related to the mass of a potential partner and negatively affected by the combined distance between them. This result based on our definition is consistent with our analysis above.

After calculation, a co-author list ranked by magnitude of acceleration can be formulated. We then pick up partner according to the sequence in that list.

Similarly, if a newly-founded company wants to build its reputation in a fast way, then wise choice of partner is critical. Our measurement tool "acceleration" will be quite useful. And if a small country wants to improve its regional or international influence in a short period, then selection of nation to cooperate is of strategic importance. It can be anticipated that USA tends to on top of lists of many countries due to both USA's single largest "mass" and its extensive cooperation with other countries in various sectors.

VIII. Sensitivity analysis

Most of the network models(including ours) are complex and abstract, and as a result the input/output relationships and many of the characteristics of the networks cannot easily be understood. A simple and efficient way to investigate these complex network models is through sensitivity analysis via simulations. Such method allows us to explore the properties of networks from an empirical view without knowing much physical law behind them. Besides, there are quite a lot uncertain phenomena exist in the nature which are difficult to discover unless the techniques of statistic are applied.

In this section, the Erdos1 network will be used to do sensitivity analysis, testing the applicability of our importance measures to this network as well as the robustness of our approaches under small disturbance. Two kinds of simulation will be conducted and the changes of input and output will be investigated. The inputs are those needed to be calculated before computing MPR(including distance, mass, and so on) and the MPR is naturally the output.

The first simulation is to randomly add disturbance into the inputs and observe the output results. The second one is gradually multiply the inputs by an arithmetic sequence which is a little bigger than 1. To gain more accuracy, 1000 sample data is taken from both types of simulations

8.1 **Distance disturbance and Mass disturbance**

In the first part, we investigate two kinds of random disturbance. One is to disturb the distance of the combined distance matrix D_c and the other is to disturb every element in the mass vector $\mathbf{m}$.

We want to know the influence of the disturbance which has -10%~10% of the original value, and hence the matlab function *rand(dim)* is used to create random numbers uniformly distributed in [0.9,1.1] and will later be multiplied by the original values to get disturbed input.

After doing so, it is of interests to know the difference between the original MPR and

the disturbed MPR. This is known the error and our measure is the mean of all the absolute value of the elements of a matrix or a vector. It is also important to observe the relationship between the input variation and output variation and therefore both them are recorded. We denote the first error as dis_error and the other as mass_error and the variation in output is simply called changes. The simulation procedure and the results are presented below:

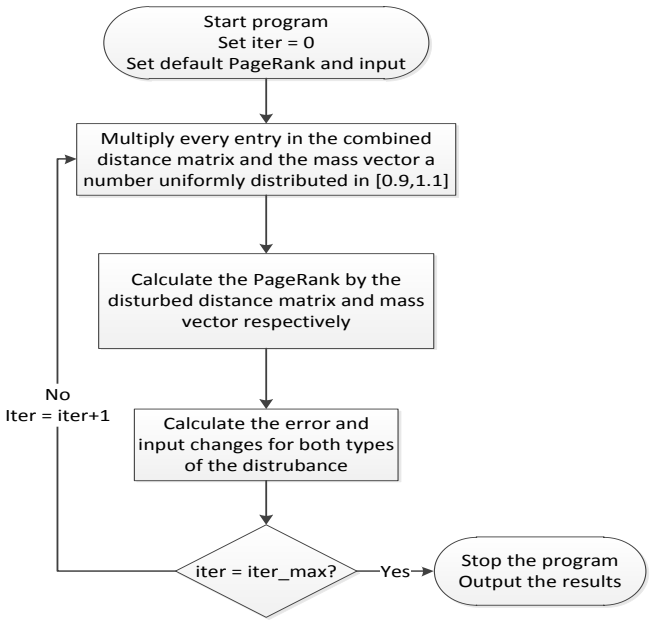

Fig10

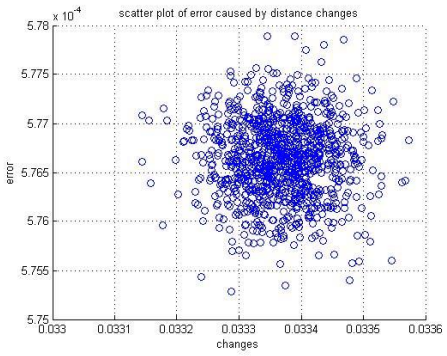

Fig11

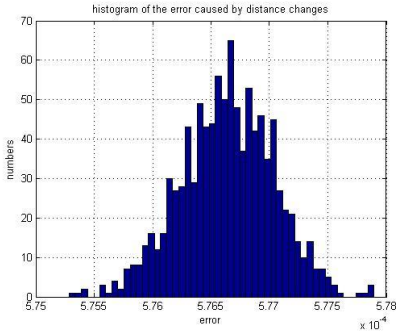

Fig12

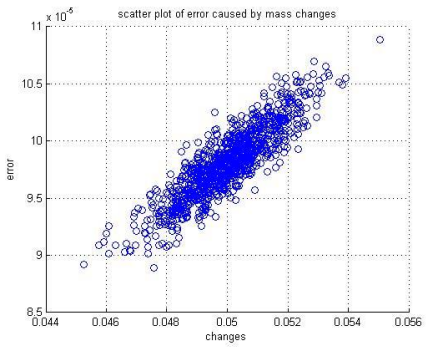

Fig13

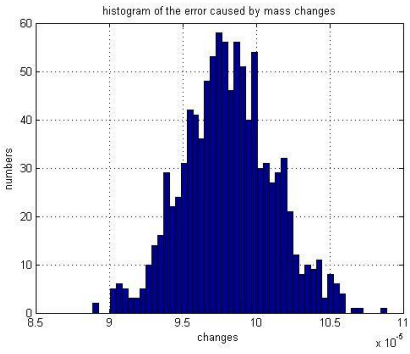

Fig14

8.2 **discussion**

From Fig (11), it is hard to say that there exists any relation between the changes and errors. One reasonable explanation is that the small disturbance in D_c seems to be have no impact on the output and hence the scatter plot appears to be like many random events. In contrast, Fig (13) shows a strong linear relationship between changes and errors. It is natural to guess if such relationship may continue for bigger variation in mass and maybe there exist some mathematical descriptions for this phenomena.

From another perspective of view, the dimension of changes and errors in Fig (13) are the same because the input and output is both a vector of same length. However, in the scenario of Fig (11), the input is a matrix and output is a vector ! And hence it's not anticipated to observe any linear or regular relationship between changes and errors. Therefore the result of Fig (11) is not so strange as we first thought.

As for Fig (12) & (14), it is obvious that both appear to be a normal distribution. This implies that under small changes, the output errors tend to be normal distributed. This result exactly accords with what we have learned in textbook. What is worthy to note is that the errors caused by mass disturbance is smaller. This may due to the completeness of D_c. Since the amount of edges are significantly greater than nodes, it is nature to see bigger errors.

IX. Strengths and Weaknesses

9.1 Strengths

The strengths of our models are mainly reflected in three aspects.

Firstly, it is quite common that actual relationship networks among entities (individuals and organizations) are not perfectly interconnected. In most cases, "isolate islands" have to be deliberately neglected in order to conduct research, which is questionable. In contrast, the introduction of concept of "virtual distance", takes potential possibility of partnership into consideration. As exemplified in Task 2 and 4, such cooperation might be realized by an "intermediary agent" like Erdos or by working for same institution like Universal. Therefore, our analysis about real world relationship is no longer restricted by predetermined arcs in the network and can be applied to a wider range of circumstances.

Secondly, generalization of law of universal gravitation in physics into network science successfully quantifies the interaction between nodes in the graph. Note that the variables defined by method of analogy like "Mass" and "Gravity" have satisfactory, if not perfect attribute in terms of interpreting network features. Desirable results obtained from our method to some extent reveal the universality and consistency of fundamental natural laws.

Thirdly, application of widely-used MPR algorithm in our analysis has two advantages. For one thing, interaction between each two nodes has been considered in the transition matrix. For another, such algorithm can accommodate dynamic characteristics of the network and a stationary state can be simulated theoretically..

9.2 Weaknesses

Our models constructed have mainly three shortcomings.

Firstly, Accumulative Citation Algorithm (ACA) used in Task 3 can only be applied to a directed tree. In other words, contribution of one node on another node should be unidirectional. Since most of the networks in reality are in the form of cooperation or interaction, the scope of application of ACA is relatively limited.

Secondly, in the first assumption, we neglect the prestige of citers to simplify our discussions in Task 3. However, it is evident that a citation by a renowned scientist greatly adds to the influence of a particular paper. Therefore, significance of a research paper is not only determined by number of cited times, but also by who refer to it. And in turn, it is logical to measure the reputation of an author by the total significance of his or her published theses. From the above analysis, we can notice that simultaneous problem exists since in academic field, prestige of scientists and significance of papers are interacted. Based on our research, an iterative algorithm [10] can be derived if a database containing information of all existing authors and papers is available.

Thirdly, during the programming, we noticed that processing speed of software Matlab was not preferable and some unexpected errors occurred at times. This fact indicates that Matlab may not be powerful enough to calculate eigenvector in a fast and accurate manner when dimension of transition matrix exceeds 1000. Some advanced computation tools may be necessary for analysis pertaining to greater networks. Otherwise, models that require less calculation workload for software should be constructed.

X. References

[1] Newman, M. E. J. "The mathematics of networks." *The new palgrave encyclopedia of economics* 2 (2008): 1-12.,

[2] Shengwei Mei et al., “SOC and Complex Networks” in *Power Grid Complexity*, Beijing: Tsinghua University Press, 2011, pp. 92-94

[3] Wasserman, Stanley. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994, pp. 220-225

[4] Cleve Moler, “Google PageRank” in *Experiments with MATLAB*, 2011 Available: http://www.mathworks.com/moler/exm/chapters/PageRank.pdf

[5] Spizzirri, Leo. "Justification and Application of Eigenvector Centrality."

[6] MatlabBGL: https://www.cs.purdue.edu/homes/dgleich/packages/matlab_bgl/

[7] American Mathematical Society: http://www.ams.org/home/page

[8] Google Scholar: http://scholar.google.com

[9] Internet Movie Database (IMDB): http://www.imdb.com/

[10] Zhou, Yan-Bo, Linyuan Lü, and Menghui Li. "Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity." *New Journal of Physics* 14.3 (2012): 033033.